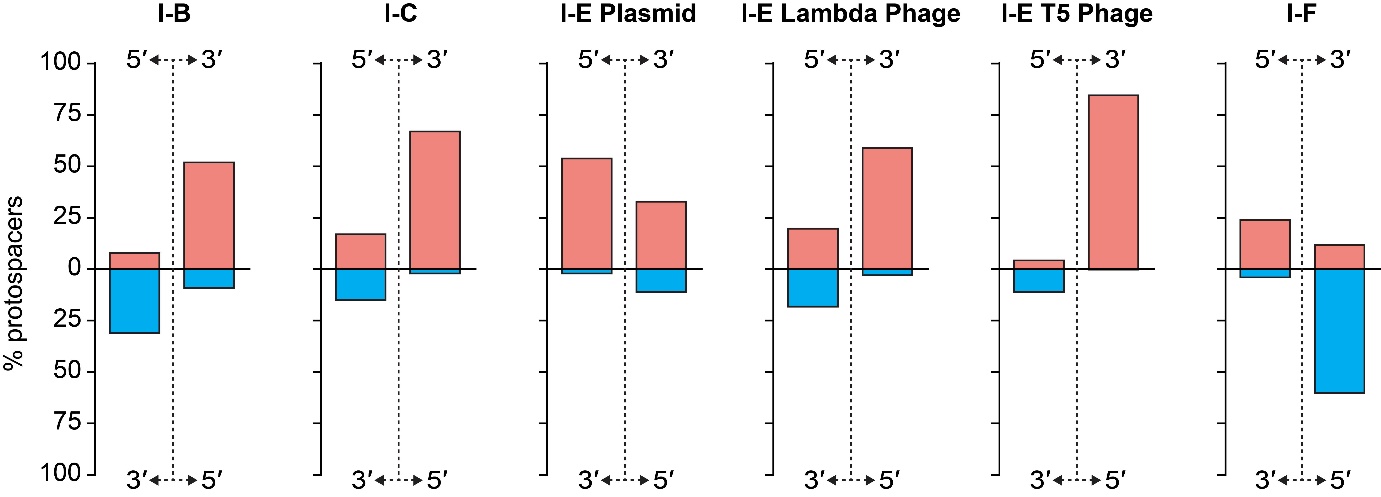
**Supplementary Information**

**Bioinformatic evidence of widespread priming in Type I and II CRISPR-Cas systems**

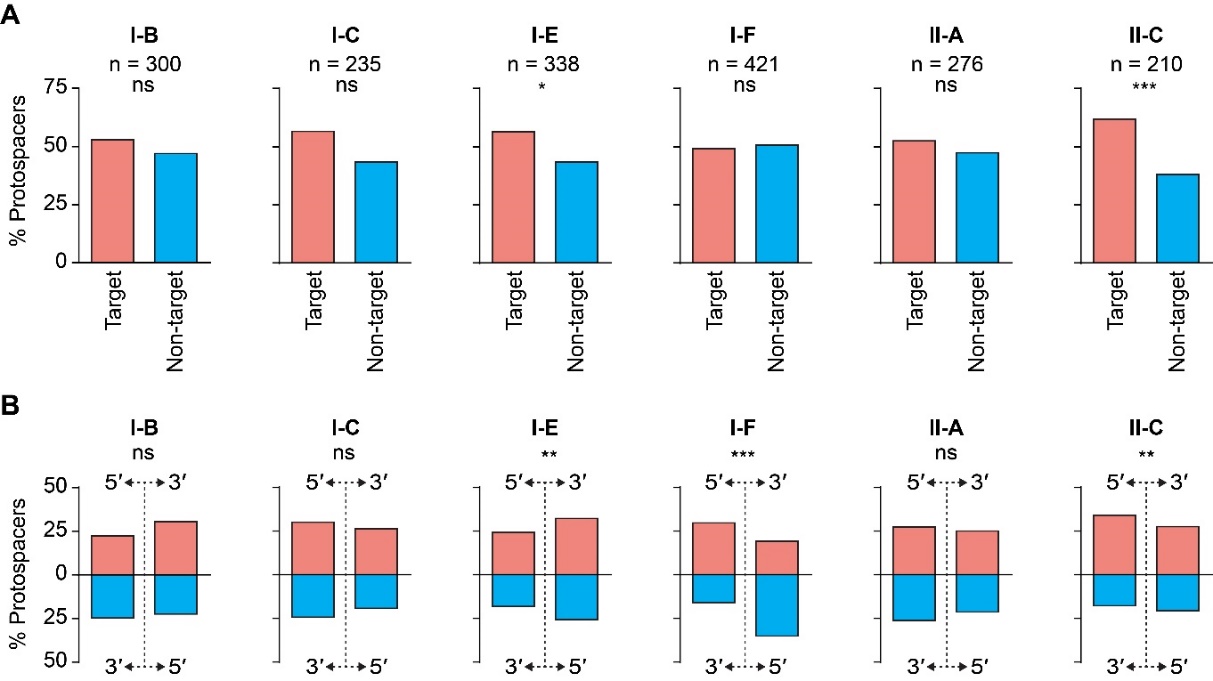
Thomas J. Nicholson1,2,#, Simon A. Jackson2,3,#, Bradley I. Croft1, Raymond H.J. Staals3,4, Peter C. Fineran2,3\*, Chris M. Brown1,2\*



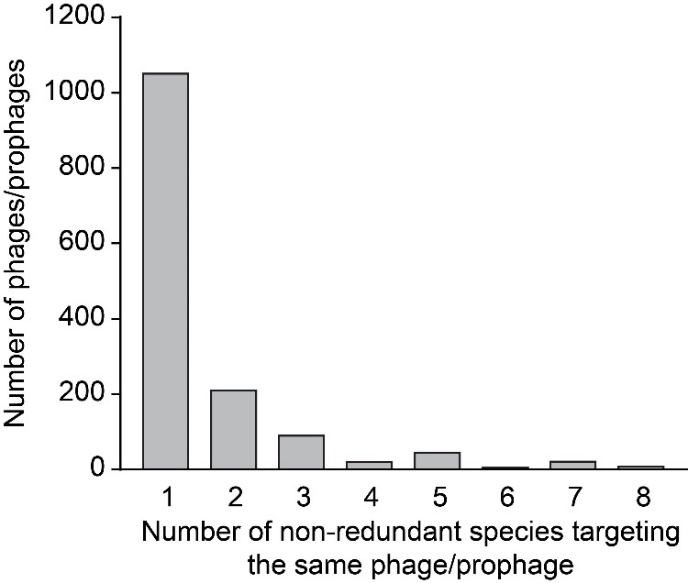
**Figure S1.** Taxonomic tree of the prokaryotic genera that contribute to the final dataset used to analyse spacer acquisition. The bars shows how many genomes from each genus contributed to the data (1-162 times; log scale) and the red boxes are showing which subtypes each genus had present in genomes.



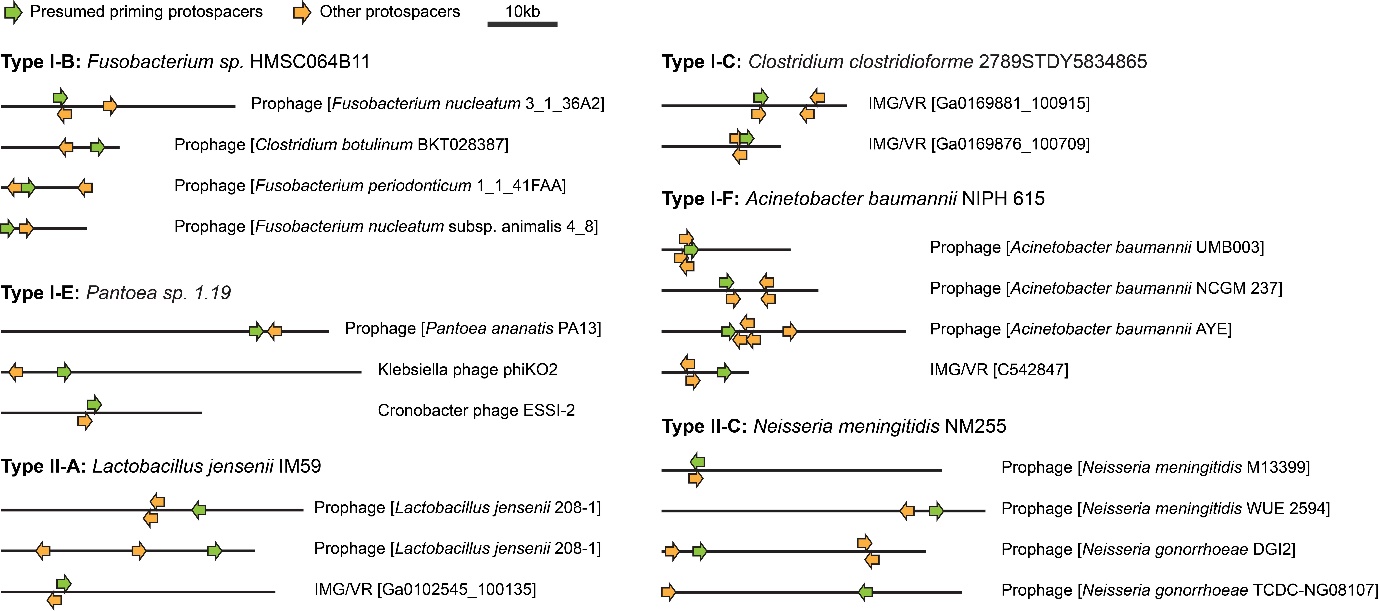
**Figure S2**. Summarised experimental data for previously characterised type I systems, showing the proportion protospacers mapping to the target and non-target strands and either 5’ or 3’ direction along these strands relative to the position of the priming protospacer. Note, that some of the corresponding studies mapped spacers, but here we compare protospacers for all systems. The experimental data are from studies of type I-B [1], I-C [2], I-E [3, 4] and I-F [5] systems.



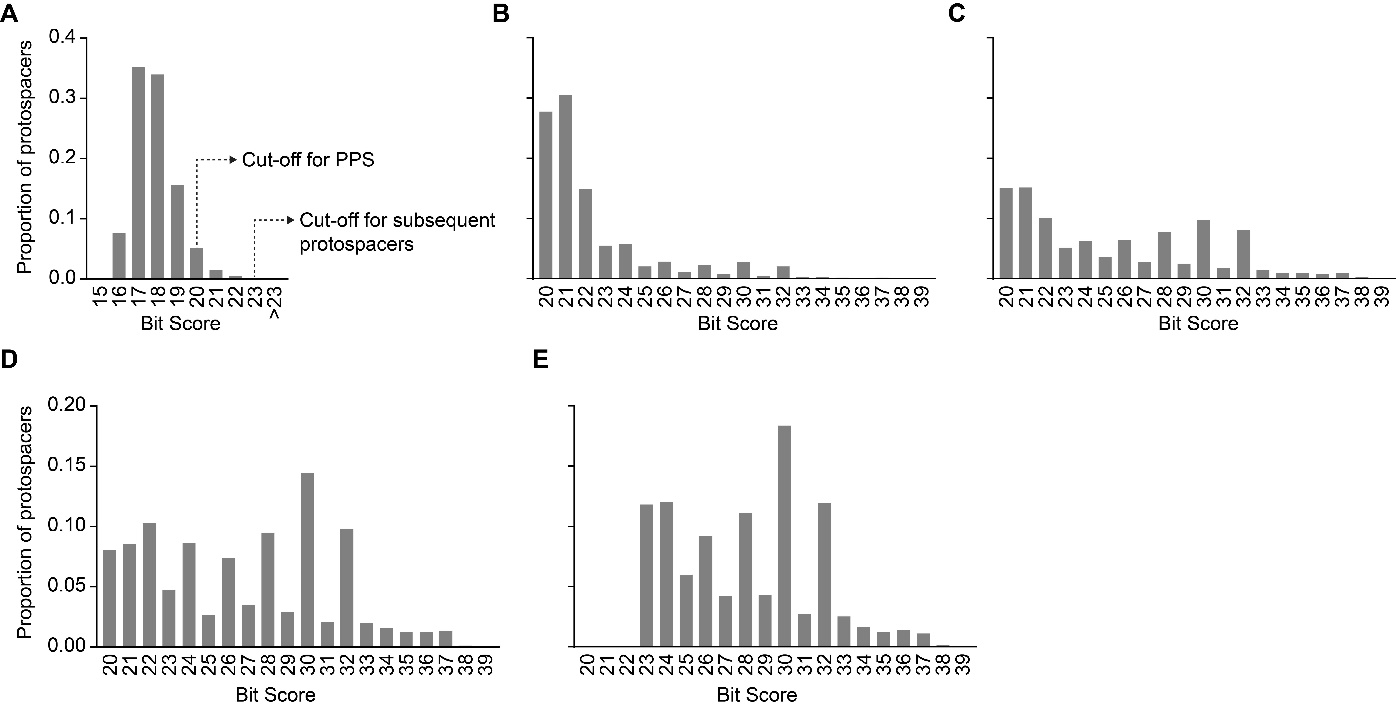
**Figure S3.** Strand and direction analyses for all data. **A)** The proportion of protospacers present on each strand across the whole genome. The significance of the differences between strands were determined using a binomial test; p >0.05 non-significant (ns), p <0.05 \*, <0.01 \*\*, <0.001 \*\*\*. **B)** Directional bias for all data, summarised by plotting the proportion of protospacers found in each quadrant (strand and direction). The significance of the distributions were determined using a multinomial test; p >0.05 non-significant (ns), p <0.05 \*, <0.01 \*\*, <0.001 \*\*\*.



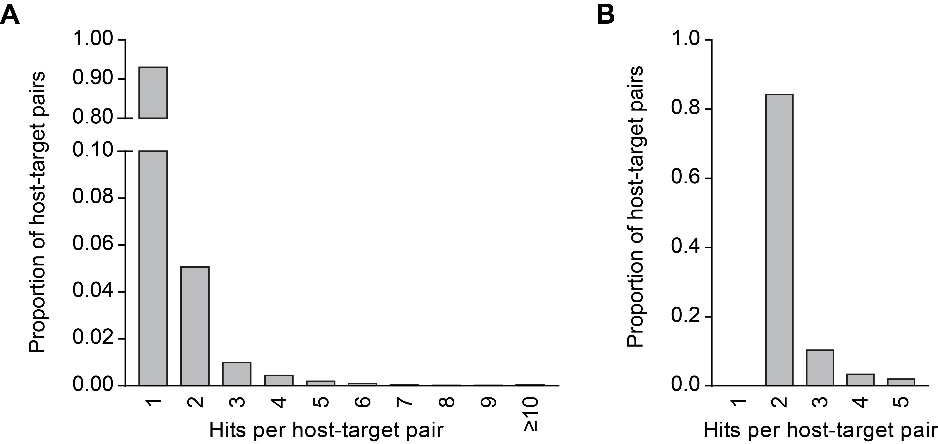
**Figure S4.** The occurrence of overlapping phage/prophage genomes between host-target pairs in the final dataset. Most phages/phages were targeted by a single representative host’s spacers (note that hosts with the same spacer/protospacer sets were merged to single representative hosts), whereas some phages/prophages were targeted by several hosts with different spacer/protospacer sets.



**Figure S5**. Examples of single hosts that target multiple phages/prophages. Note that protospacers toward either end of the displayed linear phage genome are considered clustered because most phage genomes undergo circularised states during replication. The IMG/VR accessions represent viral contigs assembled from metagenomic data [6]. In some cases the displayed host species are representative of several related hosts that possess the same spacer sets matching the specified phage.­­­­



**Figure S6.** Distributions of the spacer-protospacer matches (hits) relative to their SWIPE bitscores. **A)** The di-nucleotide shuffled dataset, with the bit score cut-off set at ≥15. The bit score cut-offs for the PPS and subsequent protospacer are as marked. **B)** All of the initial hits with the cut-off set at ≥20. The low initial cut-off score was necessary to ensure sensitive detection of putative priming protospacers. Low quality false positive hits were subsequently reduced during data processing and redundancy filtering. **C)** The bitscore distribution for matches in (B) filtered to include only host-target pairs with at least 2 hits. There is an enrichment for higher scoring data when single-hit host-target pairs are eliminated, demonstrating a reduction in likely false positive hits. **D)** The priming protospacer bitscores in the final non-redundant data set. **E)** Bitscores of the protospacers in the final non-redundant dataset that were determined to be acquired subsequent to the priming protospacers.



**Figure S7.** Spacer-protospacer matches per host-target pair. **A)** The distribution of host-target pairs relative to the number of spacer-protospacer matches (hits) within each pair for the initial SWIPE matches. The high proportion of single-hit pairs was due to the low bitscore threshold for initial matches that was necessary to detect putative priming protospacers (**Fig. S6**). **B)** The hits per host-target pair distribution for the final non-redundant dataset.

**Supplementary References**

1. Li M, Wang R, Zhao D, et al. Adaptation of the *Haloarcula hispanica* CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res. 2014.42:2483-92.

2. Rao C, Chin D, Ensminger AW. Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. RNA. 2017.23:1525-1538.

3. Savitskaya E, Semenova E, Dedkov V, et al. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. RNA Biol. 2013.10:716-25.

4. Strotskaya A, Savitskaya E, Metlitskaya A, et al. The action of *Escherichia coli* CRISPR-Cas system on lytic bacteriophages with different lifestyles and development strategies. Nucleic Acids Res. 2017.45:1946-1957.

5. Staals RH, Jackson SA, Biswas A, et al. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. Nat Commun. 2016.7:12853.

6. Paez-Espino D, Chen IA, Palaniappan K, et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. 2017.45:D457-D465.